CLAIMS

What is claimed is:

1. A method for parsing documents in query processing, said method comprising:

producing at least one index of a document written in a mark-up language;

corresponding said index to said document;

scanning said document; and

selectively skipping portions of said document based on instructions from said index.

2. The method of claim 1, wherein said mark-up language comprises any of HTML and XML.

3. The method of claim 1, wherein the skipped portions of said document comprise portions irrelevant to said query.

4. The method of claim 1, wherein said index comprises a plurality of elements representing textual categories of said query.

5. The method of claim 4, wherein said instructions match said elements to said query.

6. The method of claim 4, wherein if said elements do not match said query, then said parser uses said index to skip the portions of the document corresponding to unmatched elements.

7. The method of claim 4, wherein said each of said elements corresponds to a position in said document.

8. The method of claim 7, wherein said position comprises an end position.

9. The method of claim 8, wherein said index uses said end position as a marker for determining where to resume scanning said document upon skipping said portions of said document.

10. The method of claim 9, wherein said elements comprise sub-elements representing textual sub-categories of said query.

11. The method of claim 10, wherein said sub-elements updates said position in said document upon skipping said portions of said document and resuming scanning of said document.

12. The method of claim 4, further comprising saving said textual categories into a buffer.

13. A system for parsing documents in query processing, said system comprising:

at least one index corresponding to a document written in a mark-up language;

a processor operable for scanning said document; and

a parser operable for selectively skipping portions of said document based on instructions from said index.

14. The system of claim 13, wherein said mark-up language comprises any of HTML and XML.

15. The system of claim 13, wherein the skipped portions of said document comprise portions irrelevant to said query.

16. The system of claim 13, wherein said index comprises a plurality of elements representing categories of said query.

17. The system of claim 16, wherein said instructions match said elements to said query.

18. The system of claim 16, wherein if said elements do not match said query, then said parser uses said index to skip the portions of the document corresponding to unmatched elements.

19. The system of claim 16, wherein said each of said elements corresponds to a position in said document.

20. The system of claim 19, wherein said position comprises an end position.

21. The system of claim 20, wherein said index uses said end position as a marker for providing said processor instructions for determining where to resume scanning said document upon said parser skipping said portions of said document.

22. The system of claim 21, wherein said elements comprise sub-elements representing sub-categories of said query.

23. The system of claim 22, wherein said sub-elements provide instructions for updating said position in said document upon said parser skipping said portions of said document and said processor resuming scanning of said document.

24. The system of claim 16, further comprising a buffer operable for saving said textual categories.

25. A program storage device readable by computer, tangibly embodying a program of instructions executable by said computer to perform a method for parsing documents in query processing, said method comprising:

producing at least one index of a document written in a mark-up language;

corresponding said index to said document;

scanning said document; and

selectively skipping portions of said document based on instructions from said index.

26. The program storage device of claim 25, wherein said mark-up language comprises any of HTML and XML.

27. The program storage device of claim 25, wherein the skipped portions of said document comprise portions irrelevant to said query.

28. The program storage device of claim 25, wherein said index comprises a plurality of elements representing textual categories of said query.

29. The program storage device of claim 28, wherein said instructions match said elements to said query.

30. The program storage device of claim 28, wherein if said elements do not match said query, then said parser uses said index to skip the portions of the document corresponding to unmatched elements.

31. The program storage device of claim 28, wherein said each of said elements corresponds to a position in said document.

32. The program storage device of claim 31, wherein said position comprises an end position.

33. The program storage device of claim 32, wherein said index uses said end position as a marker for determining where to resume scanning said document upon skipping said portions of said document.

34. The program storage device of claim 33, wherein said elements comprise sub-elements representing textual sub-categories of said query.

35. The program storage device of claim 34, wherein said sub-elements updates said position in said document upon skipping said portions of said document and resuming scanning of said document.

36. The program storage device of claim 28, further comprising saving said textual categories into a buffer.

37. A system for efficiently parsing documents in query processing, said system comprising:

means for producing at least one index of a document written in a mark-up language;

means for corresponding said index to said document;

means for scanning said document; and

means for selectively skipping portions of said document based on instructions from said index.